

Multiancestry transferability of a polygenic risk score for diverticulitis

Thomas E Ueland ,¹ Jonathan D Mosley,^{2,3} Christopher Neylan,⁴ John P Shelley,¹ Jamie Robinson,^{3,5} Eric R Gamazon,⁶ Lillias Maguire,⁴ Richard Peek,⁷ Alexander T Hawkins⁸

To cite: Ueland TE, Mosley JD, Neylan C, *et al.* Multiancestry transferability of a polygenic risk score for diverticulitis. *BMJ Open Gastroenterol* 2024;**11**:e001474. doi:10.1136/bmjgast-2024-001474

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjgast-2024-001474>).

Received 26 May 2024
Accepted 6 September 2024

ABSTRACT

Objective Polygenic risk scores (PRS) for diverticular disease must be evaluated in diverse cohorts. We sought to explore shared genetic predisposition across the phenome and to assess risk stratification in individuals genetically similar to European, African and Admixed-American reference samples.

Methods A 44-variant PRS was applied to the *All of Us* Research Program. Phenome-wide association studies (PheWAS) identified conditions linked with heightened genetic susceptibility to diverticular disease. To evaluate the PRS in risk stratification, logistic regression models for symptomatic and for severe diverticulitis were compared with base models with covariates of age, sex, body mass index, smoking and principal components. Performance was assessed using area under the receiver operating characteristic curves (AUROC) and Nagelkerke's R^2 .

Results The cohort comprised 181 719 individuals for PheWAS and 50 037 for risk modelling. PheWAS identified associations with diverticular disease, connective tissue disease and hernias. Across ancestry groups, one SD PRS increase was consistently associated with greater odds of severe (range of ORs (95% CI) 1.60 (1.27 to 2.02) to 1.86 (1.42 to 2.42)) and of symptomatic diverticulitis (95% CI) 1.27 (1.10 to 1.46) to 1.66 (1.55 to 1.79)) relative to controls. European models achieved the highest AUROC and Nagelkerke's R^2 (AUROC (95% CI) 0.78 (0.75 to 0.81); R^2 0.25). The PRS provided a maximum R^2 increase of 0.034 and modest AUROC improvement.

Conclusion Associations between a diverticular disease PRS and severe presentations persisted in diverse cohorts when controlling for known risk factors. Relative improvements in model performance were observed, but absolute change magnitudes were modest.

INTRODUCTION

The diverticular disease spectrum spans from asymptomatic diverticulosis to complicated diverticulitis, with more severe disease conferring a substantial burden on quality of life. While previously thought that the chances of experiencing a severe episode were related to episode frequency, the first diverticulitis episode likely poses the greatest risk of complications.¹ The lack of tools for personalised stratification is especially important in light of recent guideline changes emphasising case-by-case

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Polygenic risk scores may assist in individualizing approaches to risk stratification for heritable diseases. Loci associated with diverticular disease have been established, while investigations of diverticulitis polygenic risk scores have been performed in participants genetically similar to European reference samples.

WHAT THIS STUDY ADDS

⇒ This multibiobank study found that a polygenic risk score for diverticulitis was associated with up to 86% greater odds of severe disease, with evidence of transferability to a diverse cohort.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The use of polygenic risk scores in counselling may inform decisions about escalation of care, dietary counselling or selection for endoscopic evaluation. This work provides direction for future functional investigations with a recommendation to focus on connective tissue biology.

circumstances when considering colectomy.² Lifestyle factors are important modifiers,³ but genetic variation has been shown to be a risk factor for severe presentations.⁴

Genome-wide association studies (GWAS) have identified loci linked with diverticular disease.^{5–9} One promising approach to translate these findings into a clinical tool is through a polygenic risk score (PRS), which combines effects of genetic variants across the genome into a single score for an individual. PRSs have been explored as an adjunct for personalised screening recommendations^{10–12} and to inform discussions about risk of disease complications.⁴ While conclusions regarding clinical utility have been mixed,¹³ PRSs may bring value for diseases where existing stratification approaches are limited. Towards this end, PRSs have been developed for diverticular disease using participants genetically similar to European reference



© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Thomas E Ueland;
thomasueland09@gmail.com

samples.^{4,6,14} However, European-derived PRSs perform variably when applied to other populations due to differences in linkage disequilibrium patterns, distinct risk loci resulting from varying selection pressures in independent populations and sparse availability of genetic data in non-European populations.^{15–17} Evidence of transferability across populations must be established prior to considering clinical implementation.

This study aimed to evaluate a diverticular disease PRS in a diverse cohort. Using phenome-wide association studies (PheWAS) of the PRS, we identified conditions associated with heightened genetic risk for diverticular disease. To assess performance in risk stratification, the PRS was included as a covariate alongside demographic and clinical risk factors in models predicting diverticulitis outcomes.

METHODS

Study population and phenotyping

Our primary analysis used the *All of Us* Research Program while a secondary analysis used Vanderbilt University Medical Center's biobank (BioVU). The *All of Us* Research Program is a longitudinal prospective cohort study combining electronic medical records with genetic information for 413 000 participants.¹⁸ BioVU is a genomic repository associated with a deidentified database of clinical records from a tertiary care centre.¹⁹ Adult participants age 18–90 years with available short read whole-genome sequencing data and electronic medical records were eligible for inclusion.

Phenotyping for the PheWAS analysis used phecodes for diverticular disease, which were mapped from International Classification of Disease, 10th revision (ICD-10) codes. Since positive predictive values range from 0.67 to 0.92 when identifying complications via ICD-10 codes alone,^{20, 21} we adopted a validated rule-based phenotyping algorithm for diverticulitis complications in the risk stratification portion of the analysis. This algorithm combined diagnostic codes, procedural codes, settings of care and temporal relationships between codes (online supplemental material).²¹ A person was defined as having severe diverticulitis when they had either two inpatient admissions for diverticulitis or a procedure performed for diverticulitis (colectomy or percutaneous drain). Mild diverticulitis was defined as receiving a diagnostic code for diverticulitis in the outpatient setting or with no more than one inpatient admission, while cases of asymptomatic diverticulosis were exclusively assigned diagnostic codes for diverticulosis. Participants with second-degree relatedness or greater were excluded via the auxiliary relatedness files from the *All of Us* Research Program and via PLINK2's implementation of KING's robust estimator for BioVU.²² Participants with a diagnostic code for colorectal cancer or an inflammatory bowel disease were also excluded. Covariates were extracted from structured clinical data elements including age, sex, body mass index (BMI) and ever smoking status.

Genotyping, imputation and quality control

The *All of Us* Research Program Controlled Tier V7 short read whole-genome samples were sequenced on the Illumina NovaSeq 6000 system with a standardised variant and sample quality control pipeline.^{23,24} In BioVU, 91 449 individuals were directly genotyped with the Illumina Expanded Multi-Ethnic Genotyping Array followed by imputation of autosomal small nucleotide polymorphisms (SNPs) to the TOPMed version R2 reference panel using the TOPMed imputation server.²⁵ Preimputation and postimputation quality control was performed to limit analyses to high-quality SNPs and samples (online supplemental material).^{26,27} We identified samples genetically similar to gnomAD descriptors of 1000Genomes reference samples (1KG) using precomputed data in the *All of Us* Research Program and principal component analysis in BioVU.²⁸ These do not reflect distinct biologic groups but represent artificial thresholds when considering genetic similarity to reference panels.²⁹ Throughout the manuscript, we adopt EUR, AFR, and AMR abbreviations to refer to individuals with genetic similarity to the reference European (EUR), African (AFR) and Admixed American (AMR) 1KG samples, respectively. Alternative methodologies for assigning genetic similarity were explored with minimal differences in the distribution of groups (online supplemental material).

Polygenic risk score

The 44-SNP PRS used in this study was previously derived from a UK Biobank GWAS of European participants through conditional and joint analysis with validation in the Michigan Genomics Initiative.^{4,9} In this GWAS, pathway enrichment analysis suggested immunity, cell adhesion, membrane transport/signalling and intestinal motility as biological processes important to the onset of diverticular disease. In the PRS, the five variants with the largest effect sizes (rs763969618, rs115490395, rs56044859, rs4333882 and rs1802575) were associated with genes implicated in membrane transport (SLC35F3, SLC6A17) and cell adhesion (ELN, EFEMP1, PCDH10).

For each individual in our cohort, a PRS was calculated with PLINK2 by multiplying the dosage of risk-increasing alleles by their reported effect sizes, and summing across all PRS loci.⁴ Scores were standardised to a mean of zero and SD of one within each group.³⁰ As a sensitivity analysis, we calculated a genome-wide PRS comprising 1 615 623 SNPs with the Bayesian PRS-CSx, which is optimised for incorporation of multiple ancestral backgrounds.³¹ For this, we used summary statistics from the largest available EUR GWAS for diverticular disease⁶ and the pan-UK Biobank AFR GWAS.³²

Statistical analysis

PheWAS

To explore clinical diagnoses associated with genetic risk for diverticular disease, we performed PheWAS of the

PRS in the *All of Us* Research Program. We included AFR, AMR and EUR given that diverticular disease code counts were greater than 200,³³ and case assignment required a minimum of two code instances on distinct days. Each ancestry group was analysed independently in addition to a combined group with all participants included. Logistic regression models were fit across the phenome, including the PRS as a covariate alongside the first 10 genetic principal components, age and sex. In sensitivity analyses, we adjusted for BMI and excluded diverticular disease cases to evaluate whether associations were contingent on a simultaneous diverticular disease diagnosis. Analysis was performed using the PheWAS R package with a two-sided p value significance threshold after Bonferroni correction for multiple comparisons (EUR: $p < 2.83 \times 10^{-5}$; AFR: $p < 3.23 \times 10^{-5}$; AMR: $p < 3.33 \times 10^{-5}$).³⁴ Next, we evaluated whether phecodes found to have association with elevated genetic risk for diverticular disease were also associated with assignment of diverticular disease diagnoses in the medical record. For this, we included significant phecodes from PRS-PheWAS as covariates in a logistic regression model with an outcome of a diverticular disease phecode when controlling for age and sex. The most conservative Bonferroni-adjusted significance threshold among ancestry groups was adopted ($p < 2.83 \times 10^{-5}$).

Clinical modelling

The primary outcome was severe diverticulitis evaluated relative to a reference group of controls with no diverticular disease. We also considered supplemental comparisons of symptomatic (mild or severe) diverticulitis as well as changing the reference group to asymptomatic

diverticulosis. Model calibration was assessed through le Cessie-van Houwelingen-Copas-Hosmer (LVC) goodness-of-fit tests and calibration plots.³⁵ Logistic regression models were fit for outcome counts exceeding 100, a criterion met for BioVU (AFR, EUR) and the *All of Us* Research Program (AFR, AMR and EUR) individuals. Covariates included the PRS, age at inclusion date, sex, BMI and ever-smoking status. A base model excluding the PRS served as a comparison from which to assess the incremental value of the PRS. Metrics for full and base models included area under the receiver operating characteristics curve (AUROC) and Nagelkerke's R^2 adjusted for the liability scale.^{30,36} All statistical analyses were performed using R V.4.3.2, and an overview of study design is shown in figure 1.

RESULTS

PheWAS

There were 181 719 individuals meeting inclusion criteria for PheWAS. For each group, the PRS was associated with phecodes for diverticular disease (EUR: $p = 1.22 \times 10^{-78}$; AMR: $p = 6.56 \times 10^{-14}$; AFR: $p = 1.55 \times 10^{-8}$) with ORs (95% CI) ranging from 1.18 (1.12 to 1.25) to 1.28 (1.20 to 1.37) (figure 2). Other PRS-phecode associations included abdominal hernia (OR (95% CI) 1.06 (1.03 to 1.08); $p = 1.76 \times 10^{-6}$) and diaphragmatic hernia (OR (95% CI) 1.08 (1.05 to 1.12); $p = 2.51 \times 10^{-6}$) for EUR models as well as diffuse diseases of connective tissue (OR (95% CI) 0.78 (0.70 to 0.87); $p = 1.07 \times 10^{-5}$) for the AMR models. When removing cases of diverticular disease, associations with connective tissue diseases but not hernias remained (online supplemental material). The inclusion of BMI as

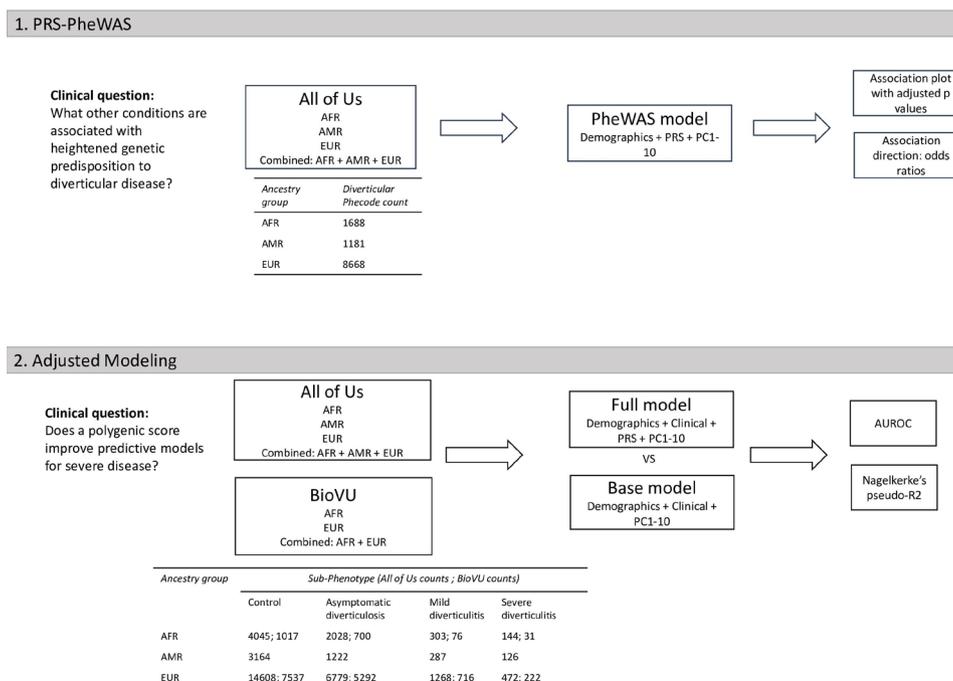


Figure 1 Overview of study design. AFR, EUR, AMR, individuals with genetic similarity to 1000Genomes African, European and Admixed American reference samples; PheWAS, phenome-wide association study; PRS, polygenic risk score.

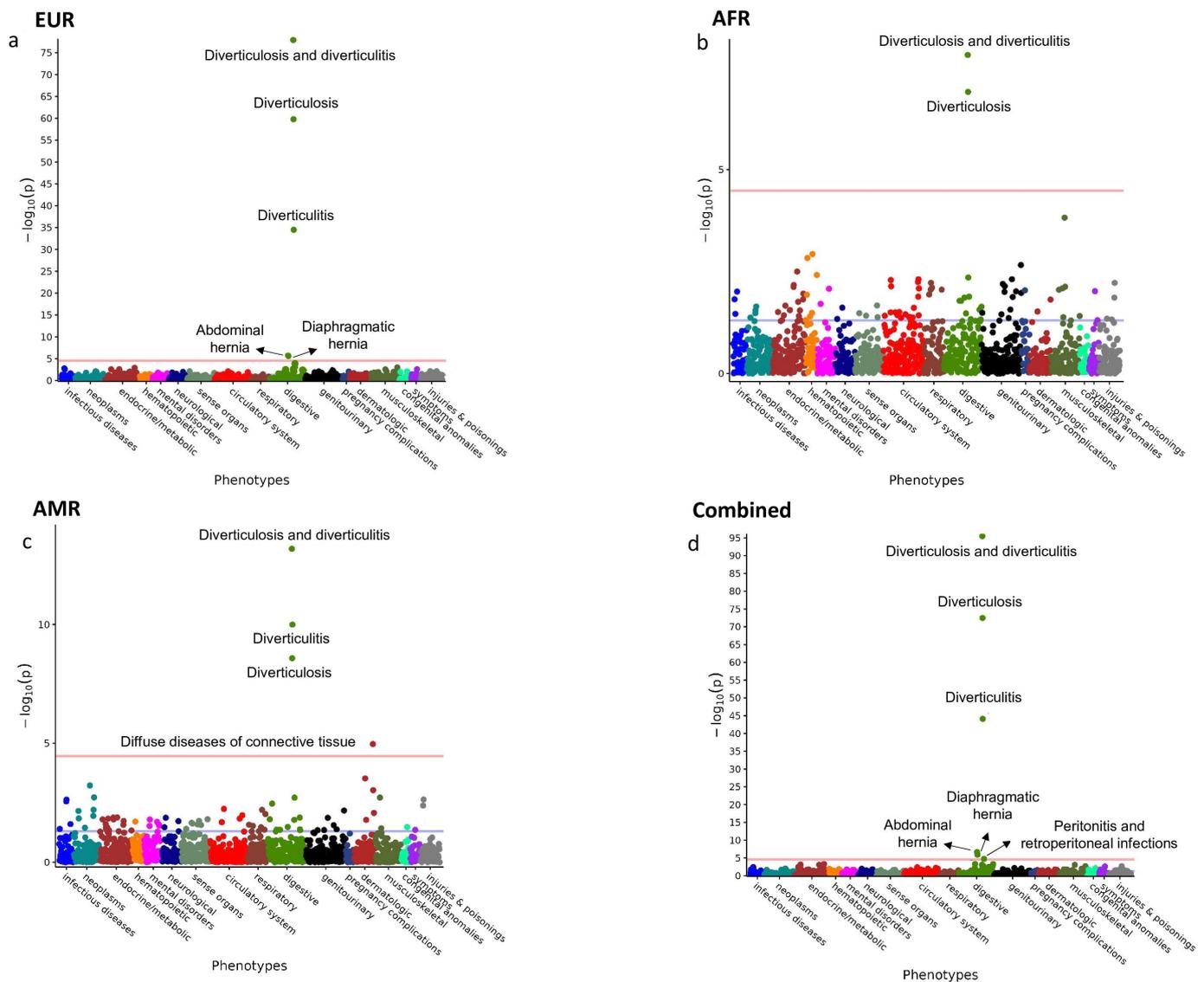


Figure 2 Phenome-wide association studies of a diverticular disease polygenic risk score in the *All of Us* Research Program. Cohorts comprised of individuals with genetic similarity to 1000Genomes reference samples in the following groups: (A) European (EUR), (B) African (AFR), (C) Admixed-American (AMR) and (D) Combined (EUR+AFR + AMR).

a covariate did not alter the significance of any findings. Among the phecodes associated with heightened genetic risk for diverticular disease, those also associated with assignment of diverticular disease codes in the medical record were inguinal hernia, diaphragmatic hernia, femoral hernia, sicca syndrome and other unspecified connective tissue disease (online supplemental material).

Clinical modelling

In the *All of Us* Research Program, the clinical modelling cohorts included EUR (n=23127), AFR (n=6520) and AMR (n=3699) individuals. Corresponding BioVU cohorts included EUR (n=13767) and AFR (n=1824) individuals. Median age was 57 (IQR 50–65) with 59% women in the *All of Us* Research Program, while median age was 56 (IQR 49–66) with 55% women in BioVU (table 1). In BioVU, severe diverticulitis comparisons were not considered for AFR due to insufficient sample size. Among all cases of diverticular disease in the *All of*

Us Research Program, the prevalence of severe diverticulitis was highest in the top quintile of polygenic risk, as compared with the lower quintiles (absolute prevalence difference, EUR: 1.3%; AMR: 2.5%; AFR: 1.6%) (online supplemental material). Models were adequately calibrated in the *All of Us* Research Program, but poorly calibrated models were found for BioVU symptomatic versus asymptomatic disease (LVC test p=0.02) and symptomatic disease versus control (LVC test p<0.01) comparisons (online supplemental material).

Covariate associations

The PRS was associated with greater odds of severe diverticulitis across groups (*All of Us* Research Program OR per SD increase (95% CI), EUR 1.65 (1.45 to 1.88); AFR 1.6 (1.27 to 2.02); AMR 1.86 (1.42 to 2.42)). This pattern persisted when broadening inclusion to symptomatic diverticulitis (*All of Us* Research Program EUR (OR (95% CI) 1.66 (1.55

Table 1 Cohort descriptions

Characteristic	All of Us=34 446*	BioVU=15 591*
Age, years	57.0 (50.0, 65.0)	56.0 (49.0, 66.0)
Female	20240 (59%)	8561 (55%)
Ancestry group		
AFR	6520 (19%)	1824 (12%)
AMR	4799 (14%)	
EUR	23 127 (67%)	13 767 (88%)
BMI (kg/m ²)	29.1 (25.3, 34.1)	27.8 (23.9, 32.6)
Ever smoker	16246 (47%)	3926 (25%)
Diverticular phenotype		
Control	21 817 (63%)	8554 (55%)
Asymptomatic	10 029 (29%)	5992 (38%)
Mild	1858 (5.4%)	792 (5.1%)
Severe	742 (2.2%)	253 (1.6%)

*Median (IQR); n (%).
AFR, African; AMR, Admixed-American; BMI, body mass index; EUR, European.

to 1.79)), AMR (OR (95% CI) 1.63 (1.39 to 1.9)) and AFR (OR (95% CI) 1.27 (1.1 to 1.46)) (table 2). When considering a reference group of asymptomatic diverticulosis, the PRS persisted as a positive predictor in EUR and AMR but not AFR models (online supplemental material).

Performance metrics

In the *All of Us* Research Program, the AUROC (95% CI) in the EUR full model for severe diverticulitis was 0.72 (0.70 to 0.75) (base 0.70 (0.68 to 0.72), difference=0.02 (0.01 to 0.03)). The AUROC (95% CI) for models of severe diverticulitis were comparatively smaller in AFR (base 0.68 (0.63 to 0.73); full 0.7 (0.66 to 0.74), difference=0.02 (0.00 to 0.04)) but not AMR models (base 0.69 (0.65 to 0.74); full 0.73 (0.68 to 0.78), difference=0.03 (0.01 to 0.06)). The trend of performance decline from EUR to AFR models was mirrored in the BioVU models for symptomatic diverticulitis (EUR base 0.75 (0.73 to 0.76); full 0.77 (0.75 to 0.78), difference=0.02 (0.01 to 0.03). AFR base 0.70 (0.65 to 0.75); full 0.72 (0.67 to 0.77), difference=0.01 (0.00 to 0.03)). The largest AUROC was observed in the EUR full model for severe diverticulitis in BioVU (0.78 (0.75 to 0.81)) (online supplemental material).

In models for severe diverticulitis, the largest Nagelkerke's pseudo-R² was achieved in the EUR models (base 0.086; full 0.106) relative to corresponding AMR (base 0.077; full 0.105) or AFR (base 0.066; full 0.085) models (figure 3, online supplemental material). Similar trends were observed with the PRS-CSx score (online supplemental material).

DISCUSSION

This multibiobank study found that positive associations between a PRS and diverticulitis persisted in AMR and AFR models with both PheWAS and risk stratification for severe disease. There were relative increases in model performance attributable to the PRS, but the magnitude of absolute change in discrimination was modest. With an optimistic transferability, the findings of this study suggest that genomic associations with diverticulitis severity subtypes may bring value to diverse clinical cohorts.

Diverticular disease is an encouraging candidate for clinical applications of genomics given a paucity of stratification approaches and estimates of heritability as high as 53%.³⁷ Three PRSs have been systematically investigated.^{4 6 14} De Roo *et al*⁴ derived a 44-SNP PRS from a European UK Biobank discovery population using conditional and joint association analysis.⁴ Wu *et al*⁶ generated a PRS using SBayesR from their meta-analysis of 724 372 European participants in the UK Biobank, FinnGen and BioVU with subsequent validation in CARTaGENE participants.⁶ Schaeffer *et al*¹⁴s 373-SNP and 851-SNP PRSs originated from application of PRSice-2 to summary statistics from UK Biobank European participants followed by validation in Geisinger MyCode participants.¹⁴ While methodologies and cohorts varied, all reported positive findings with respect to PRS validation in European individuals.

Our study's PRS-PheWAS provided initial support for validity of a diverticular disease PRS across diverse ancestral backgrounds. Diverticular phecodes were consistently associated with the PRS, with the magnitude of association largest in AMR models. We identified conditions associated with heightened genetic risk for diverticular disease, including abdominal hernias and connective tissue diseases. The disappearance of hernia phecode significance when excluding diverticular cases suggests that this association was more likely driven by a concomitant diagnosis of diverticular disease rather than true pleiotropy. Plausible explanations include incidental diverticulosis identified after an abdominal imaging study obtained for hernias, or incisional hernias occurring after an abdominal operation for diverticulitis. Time-censored PheWAS would allow for further evaluation of these hypotheses.¹² We also found that genetic susceptibility to diverticular disease was associated with lower odds of the parent phecode for diffuse connective tissue diseases, an unexpected result given that functional investigations implicate connective tissue biology in the onset of diverticular disease.^{5-7 9} It is also notable that this association was present in one subgroup (AMR) but not others (EUR, AFR). There may be differential contributions from the diagnosis codes encompassed by the parent phecode for diffuse diseases of connective tissue (709). Overall, our findings support continued focus on connective tissue biology for future investigations of pathogenesis.

Table 2 Adjusted model effect sizes

	Severe diverticulitis (OR (95% CI))*†				Symptomatic diverticulitis (OR (95% CI))*†			
	EUR	AFR	AMR	Combined	EUR	AFR	AMR	Combined
All of Us Research Program								
PRS‡	1.65 (1.45 to 1.88)	1.60 (1.27 to 2.02)	1.86 (1.42 to 2.42)	1.71 (1.54 to 1.91)	1.66 (1.55 to 1.79)	1.27 (1.10 to 1.46)	1.63 (1.39 to 1.90)	1.62 (1.52 to 1.72)
Age	2.56 (2.22 to 2.96)	1.63 (1.30 to 2.05)	1.97 (1.50 to 2.59)	2.23 (2.00 to 2.49)	2.56 (2.37 to 2.77)	1.74 (1.52 to 1.98)	1.83 (1.57 to 2.12)	2.23 (2.10 to 2.37)
Female	1.02 (0.84 to 1.24)	1.98 (1.3 to 3.02)	1.07 (0.70 to 1.62)	1.13 (0.96 to 1.32)	1.12 (1.01 to 1.25)	1.76 (1.38 to 2.25)	1.11 (0.87 to 1.41)	1.19 (1.09 to 1.30)
BMI	1.42 (1.28 to 1.58)	1.25 (1.03 to 1.51)	1.45 (1.19 to 1.79)	1.4 (1.29 to 1.52)	1.44 (1.36 to 1.53)	1.37 (1.22 to 1.54)	1.42 (1.25 to 1.60)	1.44 (1.37 to 1.52)
Ever smoker	1.27 (1.05 to 1.54)	1.78 (1.23 to 2.57)	1.21 (0.82 to 1.78)	1.35 (1.15 to 1.57)	1.28 (1.15 to 1.43)	1.68 (1.35 to 2.08)	1.14 (0.91 to 1.43)	1.32 (1.21 to 1.44)
BioVU								
PRS‡	1.66 (1.38 to 1.98)	NA\$	NA\$	1.69 (1.42 to 2.00)	1.67 (1.51 to 1.84)	1.43 (1.08 to 1.89)	NA\$	1.66 (1.51 to 1.82)
Age	2.94 (2.42 to 3.58)	NA\$	NA\$	2.79 (2.33 to 3.34)	3.01 (2.72 to 3.33)	1.66 (1.30 to 2.12)	NA\$	2.83 (2.57 to 3.11)
Female	0.89 (0.67 to 1.17)	NA\$	NA\$	0.98 (0.76 to 1.27)	1.01 (0.87 to 1.16)	1.54 (0.98 to 2.40)	NA\$	1.05 (0.91 to 1.20)
BMI	1.33 (1.13 to 1.56)	NA\$	NA\$	1.34 (1.16 to 1.57)	1.30 (1.19 to 1.42)	1.28 (0.98 to 1.66)	NA\$	1.30 (1.20 to 1.42)
Ever smoker	2.01 (1.51 to 2.66)	NA\$	NA\$	2.07 (1.59 to 2.7)	1.34 (1.15 to 1.58)	2.32 (1.50 to 3.59)	NA\$	1.43 (1.23 to 1.66)

*Outcome comparisons relative to controls of no diverticular disease.
†Additional covariates included in the model but not shown: principal components 1–10.
‡The PRS was considered as a continuous variable, and ORs are per one SD increase in the PRS.
\$NA: Not applicable due to insufficient sample size (<100) for comparisons.
AFR, African; AMR, Admixed American; BMI, body mass index; EUR, European; PRS, polygenic risk score.

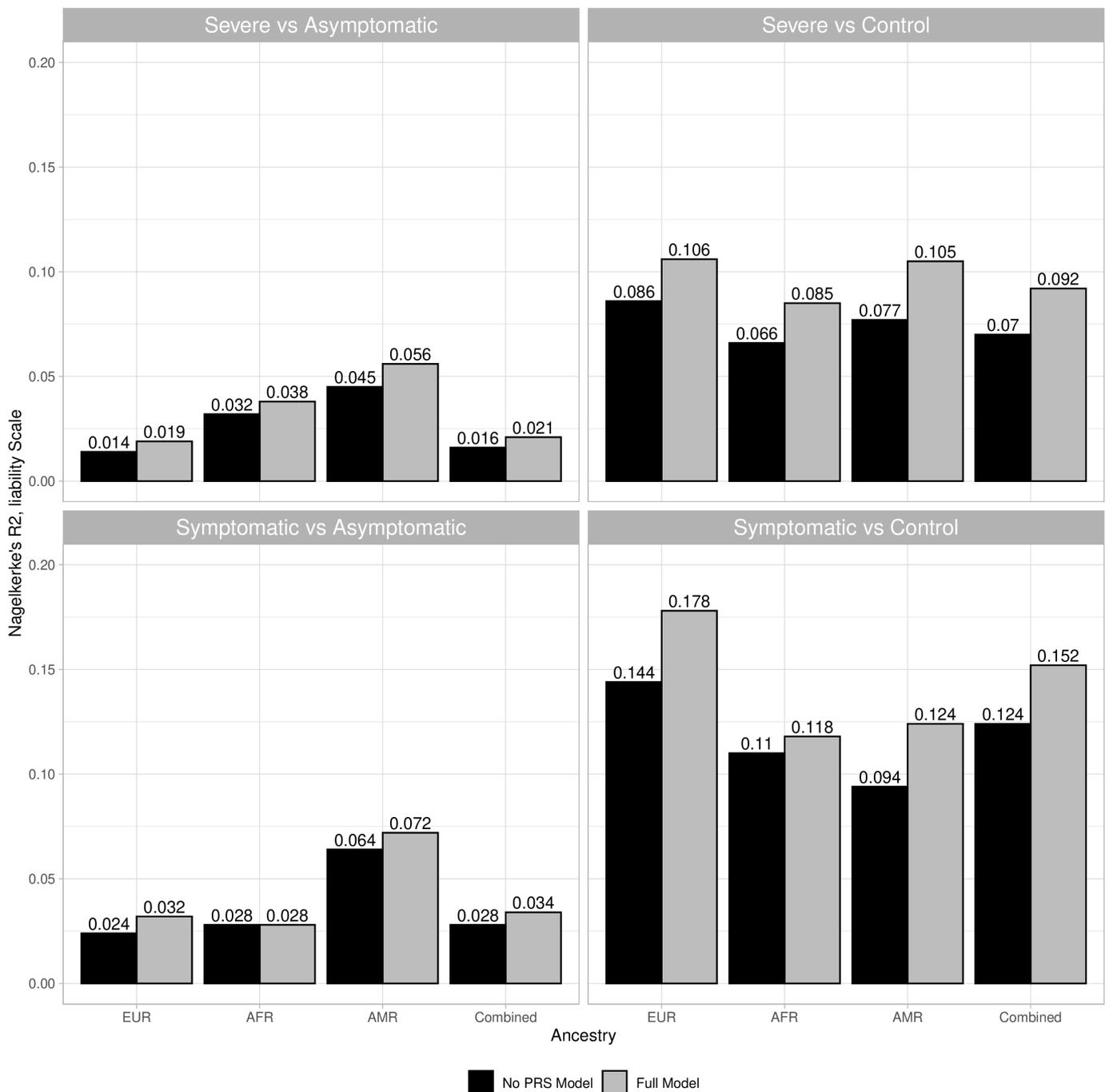


Figure 3 Nagelkerke's R^2 in the All of Us Research Program. Covariates in the base model include age, sex, the first 10 genetic principal components, body mass index and any smoking status. The full model additionally includes the polygenic risk score. The Nagelkerke's R^2 was converted to the liability scale as recommended by Lee *et al.*³⁶ AFR, EUR, AMR: individuals with genetic similarity to 1000Genomes African, European and Admixed American reference samples. Combined group includes AFR + EUR + AMR participants.

Differential performance of PRSs across ancestral backgrounds is an important consideration given the diversity encountered in many clinical practices. In our study's adjusted models, a positive association between the PRS and severe diverticulitis was found in EUR models, which persisted for AMR and AFR models. This trend of transferability suggests that diverticular disease loci from EUR discovery populations may retain predictive ability in other populations. However,

absolute improvements in AUROC and Nagelkerke's R^2 were modest in all models. The performance of European PRSs often degrades when applied to other ancestries due to differences in allele frequencies, correlation patterns between SNPs or selection pressures.^{38–40} On the contrary, some have reported similar levels of performance in other diseases,⁴¹ while a prior study of diverticular disease identified correlation between effect sizes of individual SNPs across ancestry

groups.⁶ Improving multi-ancestry applications of PRSs requires attention to practices of data collection, data analysis and return of information to participants. First, it is critical to improve the availability of genomic data in non-European populations, recently modelled through efforts from the *All of Us* Research Program and the Global Biobank Meta Analysis Initiative.^{18 42} For PRSs, the importance of adequately powered discovery GWASs in diverse populations is illustrated by our PRS-CSx sensitivity analysis. Even when adopting this PRS calculation method that is designed specifically for multi-ancestry cohorts, drastic differences between discovery GWAS sample sizes (EUR n=56 355 cases; AFR n=298 cases) likely contributed to the observed minimal improvement in performance. Second, continued work is needed in methods development specifically for cross-ancestry derivation and evaluation of PRSs.^{31 43 44} Finally, collaboration between clinicians, patients and genomic researchers is required to optimise the ways in which genomic risk information should be returned to participants. Recent examples from the eMERGE Network provide practical guidance towards this end.^{45 46}

While unlikely that a diverticular disease PRS would dramatically alter existing stratification approaches, the future potential for such a tool is promising. The degree of variance attributable to the PRS in our study did not approach existing estimates of SNP-based heritability (11%), potentially due to methodology in SNP selection and effect size estimation for the PRS as well as phenotypic differences between discovery and validation populations.⁶ For example, existing GWASs have identified cohorts with any diverticular disease using diagnostic codes, but validation studies that evaluate clinical impact narrow the focus to symptomatic, severe or recurrent diverticulitis.^{4 14} Future GWAS of severe or recurrent diverticulitis might shed light on variants specifically linked with disease progression. Furthermore, estimates of broad-sense heritability (41%) suggest possible roles for rare variants or non-additive genetic effects in diverticular disease.⁶ Additional investigation is needed to uncover the underlying biological mechanisms responsible for observed genetic associations.

A well-performing diverticular disease PRS would bring clinical value. For one, it may inform discussions about elective colectomy for patients with prior severe or recurrent diverticulitis. A discrete choice analysis of colorectal surgeons has already demonstrated substantial receptivity to a genetic risk tool in this scenario.⁴ In addition, with the onset of endoscopy-based predictive models,^{47 48} a diverticulitis PRS may be considered during initial patient evaluation to assist with targeted selection for further workup. Finally, there is evidence that awareness from the return of personalised genomic profiles motivates positive behavioural change, which would be an important addition to diverticulitis care where lifestyle factors are strong risk modifiers.^{49 50}

There are a number of limitations to this study. For both the *All of Us* Research Program and BioVU, we were unable to investigate participants similar to South Asian, East Asian or Middle Eastern reference samples due to sample size. The PRS used in this study originated from European discovery populations, and future matching of ancestry groups between discovery and validation populations may lead to different conclusions. We did not include diet and exercise in adjusted clinical models due to availability in the electronic medical records, and there are known limitations associated with using BMI as a proxy for central adiposity as well as ICD codes for smoking status. In addition, we did not include rare variants or gene by environment interactions though these are critical areas for future study.

CONCLUSIONS

Associations between a diverticular disease PRS and severe presentations persisted in diverse cohorts when controlling for known demographic and clinical risk factors. Relative improvements in model performance were observed, but magnitudes of absolute change were modest.

Author affiliations

¹Vanderbilt University School of Medicine, Nashville, Tennessee, USA

²Department of Internal Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

³Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA

⁴Department of Surgery, University of Pennsylvania, Philadelphia, Pennsylvania, USA

⁵Department of Pediatric Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA

⁶Department of Medicine, Division of Genetic Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA

⁷Division of Gastroenterology, Hepatology, and Nutrition, Vanderbilt University Medical Center, Nashville, Tennessee, USA

⁸Division of General Surgery, Section of Colon & Rectal Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Acknowledgements We would like to acknowledge Hannah Polikowsky, Lauren Petty and the Below Lab for expertise in genetic ancestry methodologies and imputation of microarray data. We also gratefully acknowledge All of Us and BioVU participants for their contributions, without whom this research would not have been possible. We thank the National Institutes of Health's All of Us Research Programme for making available the participant data examined in this study.

Contributors TEU: conceptualisation, data curation, formal analysis, funding acquisition, investigation, methodology, software, validation, visualisation, writing—original; JDM: conceptualisation, funding acquisition, methodology, resources, visualisation, writing—review; CN: methodology, supervision, writing—review; JPS: methodology, resources, software, supervision, writing—review; JR, ERG, LM: conceptualisation, funding acquisition, methodology, resources, writing—review; RP: funding acquisition, project administration, resources, writing—review; ATH: conceptualisation, funding acquisition, project administration, resources, supervision, visualisation, writing—review, guarantor. All authors have reviewed and approved the final draft submitted.

Funding TEU and RP work were supported by National Institutes of Health award number T32DK007673. ATH' work on this manuscript was supported by the National Institute of Diabetes and Digestive and Kidney Disease of the National Institutes of Health under award number K23DK118192. JDM's work was supported by the National Institute of General Medical Sciences under award number R01GM130791.

Disclaimer The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Vanderbilt University Medical Centre Institutional Review Board (IRB #230138). Work in All of Us was approved through a dedicated workspace using the Controlled Tier Dataset version 7, available to authorised users on the Researcher Workbench.

Provenance and peer review Not commissioned; internally peer-reviewed.

Data availability statement No data are available. Individual-level data are not available to protect the privacy of biobank participants. Analysis workspaces and code can be shared upon request to approved All of Us Researcher Workbench users.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID ID

Thomas E Ueland <http://orcid.org/0000-0001-6618-9318>

REFERENCES

- Strate LL, Morris AM. Epidemiology, Pathophysiology, and Treatment of Diverticulitis. *Gastroenterology* 2019;156:1282–98.
- Hall J, Hardiman K, Lee S, et al. The American Society of Colon and Rectal Surgeons Clinical Practice Guidelines for the Treatment of Left-Sided Colonic Diverticulitis. *Dis Colon Rectum* 2020;63:728–47.
- Talutis SD, Kuhnen FAH. Pathophysiology and Epidemiology of Diverticular Disease. *Clin Colon Rectal Surg* 2021;34:81–5.
- De Roo AC, Chen Y, Du X, et al. Polygenic Risk Prediction in Diverticulitis. *Ann Surg* 2023;277:e1262–8.
- Joo YY, Pacheco JA, Thompson WK, et al. Multi-ancestry genome- and phenome-wide association studies of diverticular disease in electronic health records with natural language processing enriched phenotyping algorithm. *PLoS One* 2023;18:e0283553.
- Wu Y, Goleva SB, Breidenbach LB, et al. 150 risk variants for diverticular disease of intestine prioritize cell types and enable polygenic prediction of disease susceptibility. *Cell Genom* 2023;3:100326.
- Schafmayer C, Harrison JW, Buch S, et al. Genome-wide association analysis of diverticular disease points towards neuromuscular, connective tissue and epithelial pathomechanisms. *Gut* 2019;68:854–65.
- Sigurdsson S, Alexandersson KF, Sulem P, et al. Sequence variants in ARHGAP15, COLQ and FAM155A associate with diverticular disease and diverticulitis. *Nat Commun* 2017;8:15789.
- Maguire LH, Handelman SK, Du X, et al. Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat Genet* 2018;50:1359–65.
- Schreurs MAC, Ramón Y Cajal T, Adank MA, et al. The benefit of adding polygenic risk scores, lifestyle factors, and breast density to family history and genetic status for breast cancer risk and surveillance classification of unaffected women from germline CHEK2 c.1100delC families. *Breast* 2024;73:103611.
- Temtem M, Mendonça MI, Gomes Serrão M, et al. Predictive improvement of adding coronary calcium score and a genetic risk score to a traditional risk model for cardiovascular event prediction. *Eur J Prev Cardiol* 2024;31:709–15.
- Hellwege JN, Dorn C, Irvin MR, et al. Predictive models for abdominal aortic aneurysms using polygenic scores and PheWAS-derived risk factors. *Pac Symp Biocomput* 2023;28:425–36.
- Kumuthini J, Zick B, Balasopoulou A, et al. The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review. *Hum Genet* 2022;141:1697–704.
- Schaeffer HD, Smelser DT, Rao HS, et al. Development of a Polygenic Risk Score to Predict Diverticulitis. *Dis Colon Rectum* 2024;67:254–63.
- Kim MS, Naidoo D, Hazra U, et al. Testing the generalizability of ancestry-specific polygenic risk scores to predict prostate cancer in sub-Saharan Africa. *Genome Biol* 2022;23:194.
- Ruan X, Huang D, Huang J, et al. Application of European-specific polygenic risk scores for predicting prostate cancer risk in different ancestry populations. *Prostate* 2023;83:30–8.
- Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015;47:979–86.
- Denny JC, Rutter JL, Goldstein DB, et al. The “All of Us” Research Program. *N Engl J Med* 2019;381:668–76.
- Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;84:362–9.
- Erichsen R, Strate L, Sørensen HT, et al. Positive predictive values of the International Classification of Disease, 10th edition diagnoses codes for diverticular disease in the Danish National Registry of Patients. *Clin Exp Gastroenterol* 2010;3:139–42.
- Ueland T, PheKB NX. *Diverticular disease severity, colonic*. Vanderbilt University School of Medicine, 2024. Available: <https://phekb.org/phenotype/diverticular-disease-severity-colonic>
- Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
- User Support. All of us genomic quality report. 2023. Available: <https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report>
- Venner E, Muzny D, Smith JD, et al. Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us Research Program. *Genome Med* 2022;14:34.
- Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
- Truong VQ, Woerner JA, Cherlin TA, et al. Quality Control Procedures for Genome-Wide Association Studies. *Curr Protoc* 2022;2:e603.
- Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 2011;Chapter 1:Unit 1.
- Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature New Biol* 2020;581:434–43.
- Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research, Board on Health Sciences Policy, Committee on Population, Health and Medicine Division, Division of Behavioral and Social Sciences and Education, National Academies of Sciences, Engineering, and Medicine. *Using population descriptors in genetics and genomics research: a new framework for an evolving field*. Washington, D.C: National Academies Press, 2023. Available: <https://www.nap.edu/catalog/26902>
- Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. *Nature New Biol* 2021;591:211–9.
- Ruan Y, Lin Y-F, Feng Y-CA, et al. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet* 2022;54:573–80.
- Pan UKBB Team. Pan UKBB summary statistics. 2023. Available: <https://pan.ukbb.broadinstitute.org/>
- Verma A, Bradford Y, Dudek S, et al. A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics* 2018;19:120.
- Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;30:2375–6.
- Hosmer DW, Hosmer T, Le Cessie S, et al. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16:965–80.
- Lee SH, Goddard ME, Wray NR, et al. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* 2012;36:214–24.
- Strate LL, Erichsen R, Baron JA, et al. Heritability and familial aggregation of diverticular disease: a population-based study of twins and siblings. *Gastroenterology* 2013;144:736–42.

- 38 Martin AR, Kanai M, Kamatani Y, *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* 2019;51:584–91.
- 39 Schaffer KR, Shi M, Shelley JP, *et al.* A Polygenic Risk Score for Prostate Cancer Risk Prediction. *JAMA Intern Med* 2023;183:386–8.
- 40 Larach DB, Lewis A, Bastarache L, *et al.* Limited clinical utility for GWAS or polygenic risk score for postoperative acute kidney injury in non-cardiac surgery in European-ancestry patients. *BMC Nephrol* 2022;23:339.
- 41 Pankratz N, Cole BR, Beutel KM, *et al.* Parkinson Disease Genetics Extended to African and Hispanic Ancestries in the VA Million Veteran Program. *Neurol Genet* 2024;10:e200110.
- 42 Zhou W, Kanai M, Wu K-HH, *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genom* 2022;2:100192.
- 43 Hoggart C, Choi SW, Garcia-González J. *bridgeprs*: a powerful trans-ancestry polygenic risk score method. *Genetics* [Preprint].
- 44 Lambert SA, Wingfield B, Gibson JT, *et al.* The Polygenic Score Catalog: new functionality and tools to enable FAIR research. *medRxiv* 2024:2024.05.29.24307783.
- 45 Lewis ACF, Chisholm RL, Connolly JJ, *et al.* Managing differential performance of polygenic risk scores across groups: Real-world experience of the eMERGE Network. *Am J Hum Genet* 2024;111:999–1005.
- 46 Sabatello M, Bakken S, Chung WK, *et al.* Return of polygenic risk scores in research: Stakeholders' views on the eMERGE-IV study. *HGG Adv* 2024;5:100281.
- 47 Tursi A, Brandimarte G, Di Mario F, *et al.* Prognostic performance of the "DICA" endoscopic classification and the "CODA" score in predicting clinical outcomes of diverticular disease: an international, multicentre, prospective cohort study. *Gut* 2022;71:1350–8.
- 48 Tursi A, Piovani D, Brandimarte G, *et al.* Diverticular Inflammation and Complication Assessment classification, CODA score and fecal calprotectin in clinical assessment of patients with diverticular disease: A decision curve analysis. *United European Gastroenterol J* 2023;11:642–53.
- 49 Muse ED, Chen S-F, Liu S, *et al.* Impact of polygenic risk communication: an observational mobile application-based coronary artery disease study. *NPJ Digit Med* 2022;5:30.
- 50 Stan DL, Kim JO, Schaid DJ, *et al.* Breast Cancer Polygenic-Risk Score Influence on Risk-Reducing Endocrine Therapy Use: Genetic Risk Estimate (GENRE) Trial 1-Year and 2-Year Follow-Up. *Cancer Prev Res (Phila)* 2024;17:77–84.